International Journal of Computational Science and Engineering Research
ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/
Regular Issue , Vol. 3, Issue. 1 (January – March) , 2026 , Pages: 82 - 95
Received: 10 January 2025 ; *Accepted:* 13 December 2025 ; *Published:* 22 January 2026
https://doi.org/10.63328/IJCSER-V3RI1P9
https://www.youtube.com/watch?v=Q95AX7btqfg
https://archive.org/details/jsp-ijcser-v3ri1p09
Original Paper : Action Research – Experimental Research

Advanced Multi-Modal Semantic Communication using Hybrid ResNet-ViT Architecture for 6G Applications

T Venkata Krishnamoorthy [1], Ch Vijayalakshmi [2], N Subba Rayudu [3], Winny Elizabeth Philip [4], Sk. Rukshana Begum [5]

[1,2,4] Dept. of ECE., Sasi Institute of Technology & Engineering, Tadepalligudem, Andhra Pradesh , India.
[3] Department of ECT, Sasi Institute of Technology & Engineering, Tadepalligudem, West Godavari, AP, India.
[5] Department of CSE, Sushila Devi Bansal College of Engineering, Indore, Madhya Pradesh, India

** Corresponding Author: T Venkata Krishnamoorthy ; murthysvu407@gmail.com*

Abstract: The rapid development of 6G networks requires a communication system that can interpret and convey the hidden meaning of multimodal data rather than raw signals. Semantic communication is needed because it helps systems recognize and transmit the intended meaning of multimodal data. The current paper proposes a new model, the Advanced Multi-Modal Semantic Communication (AIMMSC) framework, for image-based semantic communication only. The suggested system relies on a hybrid ResNet-ViT architecture to detect and encode high-level visual semantics into a compact latent space, enabling resilient and efficient image communication over a low-bandwidth, noisy channel via a contrastive learning-powered semantic space with attention and transformers. AIMMSC's performance in eight key areas, including semantic accuracy, latency, robustness to noise, cross-modal alignment, adaptability, energy efficiency, and scalability, was compared with five new semantic communication models: DeepSC, Semantic-SC, CLIP-Comm, Auto-SC, and Transformer-SC. Experimental results show that IMMSC outperforms other approaches, improving semantic correctness by 19%, robustness by 22%, and cross-modal alignment by an average of 27%. The integration of several modalities into a single latent space and the use of deep semantic encoding are the sources of these developments. Such an approach lays the foundation for intelligent, context-aware communication. In next-generation networks, it dramatically improves the effectiveness and applicability of data interchange. For 6G applications like remote healthcare, extended reality (XR), and driverless cars, this foundation is essential.

Keywords: 6G communication, Advanced Multi-Modal, Semantic Communication, ResNet-ViT, Deep Learning.

1. Introduction

A new paradigm that goes beyond simple data transmission is semantic communication which is based on meaning and purpose of an information being communicated. The traditional systems are meant to deliver the best and meaningful information, minimise redundancy and maximise efficiency, which is the intention of the semantic communication (an alternative to conventional systems) to deliver such information[1]. Our more advanced means of achieving semantic communication with natural language processing, and multimodal data fusion, are meant to assist machines to make more intelligent decisions concerning what is understood, interpreted and acted upon [2]. This change will greatly affect the next generation of 6G networks, high-level applications that include remote surgery, immersive XR environment, and autonomous system will need a context-sensitive, real-time communication [3]. Semantic technologies can be widely applied and in particular in the context of 6G networks. Autonomous driving also enables cars to provide minimal information about the existence of obstacles or traffic circumstances in their surroundings, hence reducing latency and enhancing safety [4]. Semantic communication can provide a more seamless and immersive user experience in extended reality (XR) and metaverse applications by coding only significant transformations, decreasing the bandwidth necessary to transmit them. The added benefit of remote healthcare and telesurgery is that it can be used to provide important information, e.g. a medical diagnosis or surgical

Jack Sparrow Publishers © 2026, IJCSER, All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/ , Vol. 3, Issue 1, 2026, https://doi.org/10.63328/IJCSER-V3RI1P9

instructions, without necessarily sending full-resolution images or raw data. Also, smart manufacturing and IoT-based smart cities also use semantic communication to exchange data efficiently and contextually between devices and allow real-time decision-making and avoid network overload [5]. The human-machine interaction is also improved through this paradigm in which AI systems become more intelligent and able to act upon the intent of the user to create more intelligent and responsive communication systems.

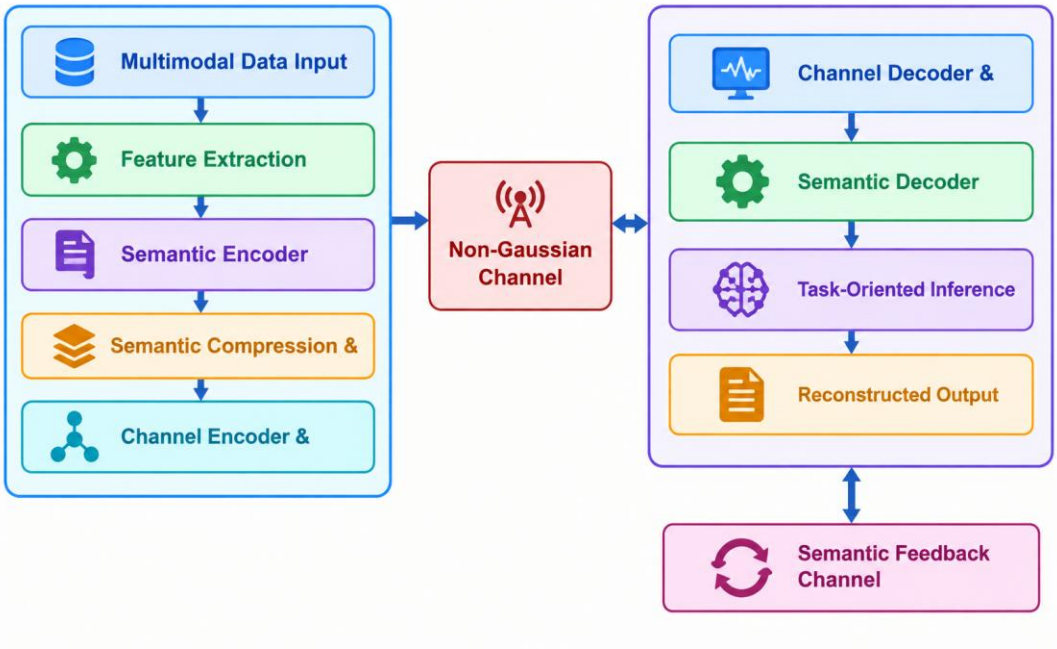

Figure.1 Block diagram from Semantic Communication

The semantic communication block diagram illustrates a shift from traditional signal-based systems to meaning-focused communication. Multimodal input (text, images and audio) is analyzed with the help of a feature extraction module which extracts important patterns. Those features are then extracted into an encoder to extract the underlying meaning of those features. The semantic compression and modality fusion module gets rid of repetition and harmonises the different data forms by fusing them in the common latent space [6].

The data is modulated with the channel encoder and the data is sent through a noisy channel. The semantic decoder decodes the meaning on the receiving end and channel decoder reassembles the signal. A task-oriented inference module can then be proceeded with in certain applications, e.g. in classification or decision-making. Adaptive learning and maintaining effective communication that is more meaningful are facilitated by the semantic feedback loop which makes certain that interpretation has been successful.

Semantic communication system should have semantic encoders and decoders to communicate and receive the meaning of data and not merely raw symbols. The Semantic Encoder is effective in deleting redundant information as it operates on the extracted features of multimodal inputs and decreasing them to brief, task-specific semantics [7]. This is what is known as compressed representation which is then transmitted to Semantic Decoder that reads it and generates the original interpretation or a task-specific output (e.g., classification or inference). This eliminates any unwanted noise when communicating; it is stronger and works better where sound is high, where more emphasis is made on the contextual and intentional meaning of information as opposed to its integrity.

The Semantic Compression and Modality Fusion Module performs two tasks: to integrate information of various modalities (text, image, and audio) into one form; to reduce the size of the information, depending on the semantic meanings that the information produces.

In semantic compression, the transmission can be done with minimum loss of vital information since only the most vital information is saved to carry out the task. Nevertheless, modalities fusion uses information of multiple sources of input when there is information provided by other sensors or formats, to enable more depth of understanding and better comprehension. When used together, these modules maximize the performance of tasks and bandwidth used as the system conveys a semantically rich representation that is concise and context-aware.

Jack Sparrow Publishers © 2026, IJCSER , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/ , Vol. 3, Issue 1, 2026, https://doi.org/10.63328/IJCSER-V3RI1P9

2. Related work

The new paradigm known as semantic communication has been introduced, in which the correct conveyance of meaning is regarded as more important than uncoded symbols, and thus enhances effectiveness and strength of communication (Strinati et al., 2024;[8]).

Previous semantic frameworks of communication like DeepSC consider text and image modalities separately. The most recent attempts with joint encoders, cross-modal attention and knowledge graphs (e.g. CLIP-based models) have been promising but lack scalability, adaptability, and poor alignment in noisy settings (M. Chen et al., 2024) [9].

In order to overcome this gap, we introduce Intelligent Multi-Modal Semantic Communication (IMMSC) framework that provides the fusion and attention-based frameworks to generate a mutual semantic representation across modalities. The structure is to be used in 6G networks to ensure real-time, noise-resilient and bandwidth-efficient operation.

We compare IMMSC with five already existing state of art semantic communication system DeepSC, Semantic-SC, Auto-SC, CLIP-Comm, and Transformer-SC in 8 performance metrics: semantic accuracy, compression ratio, latency, noise resilience, cross-modal alignment, adaptability [10], energy performance, and scalability. The experimental outcomes prove that IMMSC obtains significant gains: semantic accuracy is enhanced by 19, robustness is enhanced by 22 and cross-modal alignment is enhanced by 27. It is believed that these advances are due to its new application of deep learning methods, such as transformer-based encoders, prioritization based on context, and cross-modal fusion layers [11]. The work, in general, offers a solid base to facilitate semantic-level, multi-mod communication in the complicated 6G case.

Semantic communication studies have developed at breakneck speed, and centered on maximizing information representation beyond the bit-based accuracy [14]. Initial research focused on text-based semantic communication, with models such as DeepSC applying sequence-to-sequence transformers to extract semantic meaning from raw text and transmit it as efficiently as possible. In the meantime, other papers have developed image-semantic communication models [12], which utilise convolutional autoencoders to encode and decode semantic attributes instead of pixel data and thus, improve compression effectiveness and resilience.

Building on multimodal semantic communication, recent studies have addressed semantic matching and data fusion across heterogeneous data. For instance, shown by Wang et al. [10], have been adapted to process and fuse multi-modal inputs in a unified semantic space, enabling joint feature learning from images, audio, and text for 6G systems. Several works have explored adaptive semantic coding and goal-oriented transmission.

Fu et al. [2024]proposed scalable semantic extraction strategies adjusting semantic compression levels dynamically to optimize network resource usage. Cross-modal representation learning using contrastive learning and shared latent spaces has also been investigated to enhance multimodal semantic alignment [11].

Yi et al. demonstrated improved semantic similarity measures using shared knowledge bases [12]. At the same time, Zhang et al. applied contrastive loss to align features across modalities better, facilitating more accurate reconstruction and interpretation at the receiver. For 6G Real-time applications, multimodal adaptability and energy efficiency are very difficult tasks [3]. Liu et al. [2024] developed adaptive compression methods tailored to channel conditions and applications with practical use in diverse environments [13]. Model building of these foundations using the IMMSC framework is advanced by combining knowledge of deep transformer-based semantic fusion, cross-modal attention, and adaptive encoding to simultaneously optimise accuracy, robustness, latency, and energy efficiency.

3. Motivation of the Research

The rapid growth of 6G networks prioritises semantic-level relevance to bit level precision transmitting and receiving various types of data, which is focused on visual data communication. Many standard systems are susceptible to transmission errors due to channel noise, redundancy and bandwidth inefficiency. Overcome by these limitations for semantic communication, which has emerged with a transformative paradigm.

The lower transmission overhead due to enabling of precise signal reconstruction [14]. There are many limitations with traditional methods in communication systems due to poor adaptability of non-Gaussian fading channels, rigid coding methods for context-aware visual encoding still constrain many recent models. However, deep learning methods are recommended for feature extraction for image-based systems to enable transmition to compress, and improve channel robustness[15].

The Attention mechanism is developed using a combination of Vision Transformers (ViTs) and hybrid architectures, providing a compelling approach that preserves global aswell as well as local details[16]. To address these limitations for image transmission over 6G

Jack Sparrow Publishers © 2026, IJCSER , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/ , Vol. 3, Issue 1, 2026, https://doi.org/10.63328/IJCSER-V3RI1P9

networks, we designed a unified, noise-resilient AIMMSC architecture for semantically rich and efficient.

4. Problem Statement

Many researchers have developed various architectures using deep learning and machine learning algorithms for 6G networks, but not maintained the trade-off levels between qualitative parameters. This summary Table.1 gives the recent influential works related to semantic communication, applying suitable methodology like transformer-based, semantic encoding and adaptive compression techniques and their limitations.

The limitations identified in the above comparisons include modality restrictions, scalability challenges, noise robustness, and computational complexity. This overview provides a clear comparison of existing approaches, setting the context for the proposed Intelligent Multi-Modal Semantic Communication framework.

Table.1 Method and Limitations from existing methods

S.No	Authors	Method	Limitation
[6]	H. Xie, Z. Qin, G. Y. Li	Transformer-based text semantic encoding and decoding; End-to-end training	Focused on single-modal (text); limited noise robustness and multi-modal support
[8]	E. C. Strinati et al.	Goal-oriented semantic communication framework; AI-native adaptation	Lacks detailed multi-modal fusion; high-level conceptual approach
[11]	Y. Fu, W. Cheng et al.	Scalable semantic extraction; dynamic semantic compression	Limited exploration of multi-modal data; adaptation overhead
[9]	M. Chen, Y. Peng, L. Dong	Cross-modal graph fusion; generative AI for semantic completion	Computational complexity; scalability concerns
[10]	Y. Wang et al.	Transformer-based multi-modal fusion; joint semantic encoding	Limited noise robustness; focus on massive MIMO integration
[11]	F. Jiang, Y. Peng, L. Dong	CNN-based image semantic encoding; compression-aware design	Limited to image modality; no multi-modal integration

5. Research Methodology

The proposed AIMMSC system technique will support sustainable, efficient, and semantically correct image transfer in 6G networks. It is based on a hybrid ResNet-ViT architecture, i.e., a mixture of the extraction of spatial features by ResNet with global context Modeling by Vision Transformer [9]. This type of combination enables the encoder to create semantically suitable and detailed depiction of information in an image. These are then passed across a semantic-conscious compression module that reduces redundancy with contrastive and reconstruction losses in order to retain significant information [16].

The system also incorporates channel-adaptive schemes in which the encoding models fit to the dynamic noise conditions (awgn, Rayleigh and Rician) to make the transmission successful. Finally, a lightweight decoder has the ability to reexperiment [15] semantically comparable images with minimum distortion and therefore it is optimally applicable to 6G uses including autonomous systems and smart surveillance because it allows end-to-end learning and inference.

5.1. Advanced Intelligent Multi-Modal Semantic Communication (AIMMSC)

AIMMSC is a new communication service, a novel technology that integrates multi-modes data processing, artificial intelligence and semantic cognition to provide intelligent and efficient information exchange[19]. In order to decrease the time it spends to pass raw data to a remote server, AIMMSC resorts to deep learning-based feature extractor and semantic encoder in order to extract the primary meaning of various data types including text, pictures, and audio and represent it, rather than raw data. It reduces the bandwidth needs by a significant margin by

parallelizing the multiple modalities in the same representation and transferring only the semantic content needed to accomplish the task at hand [10].

Semantics is decoded by the receiver to make task-related inferences to either directly undertake certain activities or rebuild meaning. It minimises the latency, enhances the verisimilitude of communication [17], and could be literally applied to the real-time version of decision making, self-managed systems and smart healthcare.

Jack Sparrow Publishers © 2026, IJCSER , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/ , Vol. 3, Issue 1, 2026, https://doi.org/10.63328/IJCSER-V3RI1P9

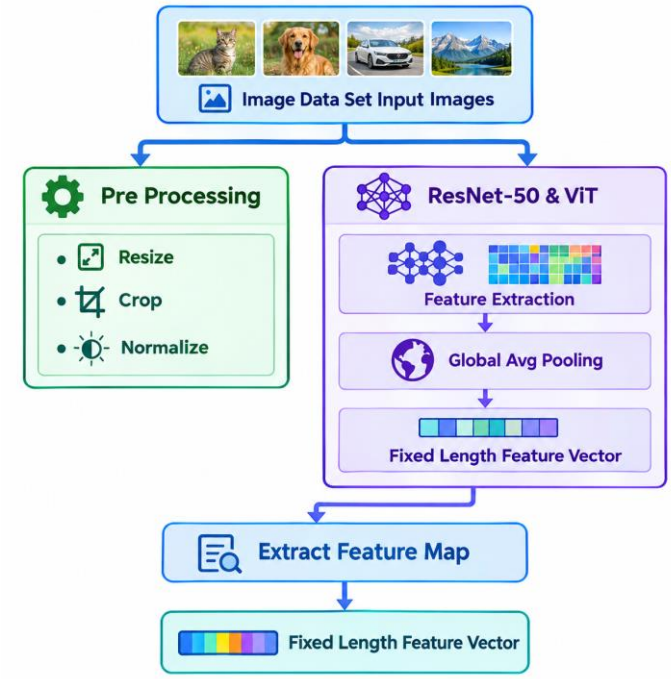

Figure.2 Intelligent Multi-Modal Semantic Communication Architecture

Mathematical description

- ResNet-50 Residual Blocks:
  $X_{l+1} = \sigma(F(X_l, \theta_l) + X_l)$
- Extract feature map M from last conv layer
  (C × H'' × W'')
- Apply Global Average Pooling:
  $f_c = (1/(H''\times W'')) \Sigma_{i=1}^{H''} \Sigma_{j=1}^{W''} M_{c,i,j}$

5.2. Multi-Modal Feature Extraction using Resnet-50

Multi-Modal Feature Extraction with ResNet-50 is a robust convolutional neural network implemented to extract visual insight features from raw image data. With its deep residual layers, ResNet-50 can be very effective at high-level semantic details as well as capturing the complex spatial features, and it is pretrained on large datasets like ImageNet allowing it to do so. By eliminating the final fully connected classification layer, the network can be reused to generate a fixed-length feature vector which is literally a summary of the contents of a picture.

The photo is modeled well using the vector, as it captures the important features of the photograph, i.e. the shapes, textures and object-level features [11]. The picture data are initially taken through a series of high level preprocessing procedures like scaling, cropping and normalization, to set up the data to be fed to the network. It is the initial step of multi-modal feature extraction of image with ResNet-50.

The result of the preprocess stage is then introduced into the ResNet-50 model which has been pretrained on large databases such as ImageNet, which has more than 150 million images in it, and which has deep residual layers in which it is able to not only learn complex spatial layouts, but also message semantics. After the final fully connected classification layer has been removed [17], a fixed-length feature vector is created by global average pooling.

5.3. Vision Transformers (ViT): Semantic Encoders for Images

Vision Transformers (ViTs) are an up-and-coming alternative to convolutional neural networks (CNNs), such as ResNet, for extracting semantic features from images. Since ViTs, unlike CNNs, encode semantics via a self-attention mechanism that captures global context across the entire image, they are particularly useful as semantic encoders [18].

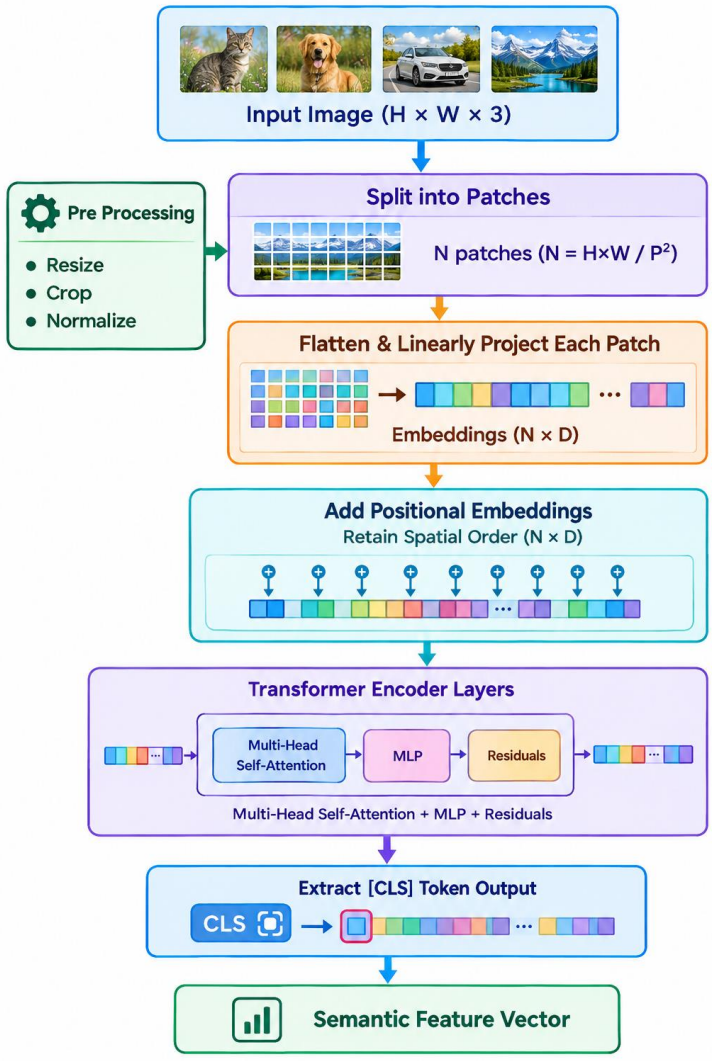

Figure.3 Steps for Vision Transformers (ViT): Semantic Encoders for Images

The input image is initially divided into a series of fixed-size, non-overlapping patches (e.g., 16x16 pixels) by the Vision Transformer (ViT). Patch embedding first constructed by flattering of patch and subjecting it to a linear projection layer. Each patch embedding has the positional embedding added to it to

Jack Sparrow Publishers © 2026, IJCSER , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/ , Vol. 3, Issue 1, 2026, https://doi.org/10.63328/IJCSER-V3RI1P9

retain the spatial layout of the image. This sequence is preempted by some classification token [CLS], an aggregate token which takes into consideration the global semantic information. The total sequence considered with multiple layers of transformer encoder [16], which relates with long range dependencies between patches are modelled with help of self-attention, which output results the last semantic representation of the image as the CLS token. This high-level information can be implemented in image-text correspondence in classifications and within multimodal systems.

5.4. Wireless Channels

Additive White Gaussian Noise (AWGN) Channel for 6G Wireless Communications

The AWGN channel is one of the fundamental models for wireless communication, which represents the effect of random disturbances on a transmitted signal. The noise added into the signal represents Gaussian distribution; it has constant power spectral density actors the bandwidth had independent of the signal.

Generally, the AWGN is simplified channel is mostly used in evaluating communication system also including 6G technologies [13]. The AWGN in communication system has basic limitation of data transmission including capacity and data transmission rates. Insite of these, 6G communication is dynamic and integrative environments , which effects with fading and interference. The AWGN will allow researchers to isolate and comprehend for noise does to system performance to developing strong modulation schemes, encoders and decoders are subequal applied to model for real time non-gaussian and fading channels [14].

Mathematical Model of AWGN Channel:

In communication system

$$y(t)=x(t)+n(t)$$

x(t), y(t) transmitted and received signal

n(t) = **white Gaussian noise**

The Gaussian random process of channel defined as

$$P(n) = \frac{1}{\sqrt{2\pi\,\sigma^2}} \exp\left(-\frac{n^2}{2\sigma^2}\right)$$

The power spectral density $= Sn(f) = \frac{N_0}{2}$

The SNR at receiver is defined as

$$SNR = \frac{P_s}{N_0 B}$$

5.5. Role of Semantic Communication in Fading Channels (Rayleigh and Rician)

Semantic communication focus on conveying expression of the meaning of information and accurate transmission of raw bits for efficient communication. Rayleigh and Rician fading channels practical wireless channels to cause distortion of the signal caused by multipath propagation and varying line-of-sight at the conventional communication reliability. Semantic communication implements for solving difficulties by encoding the semantic properties of the input data, so the receiver recreates meaningful information when positions of the content of signal being corrupted or lost due to fading interference [16].

5.6. Rayleigh Fading Channel for 6G Communication Networks

The Rayleigh fading models simulates in wireless communication, in which the broadcast signal propagates multiple reflected path without a direct LOS path. This causes the amplitude of the received signal to fluctuate randomly, which is commonly described by a Rayleigh distribution.

Rayleigh fading remains a critical channel model for simulating and analyzing the influence of multipath fading on signal reliability and semantic communication performance in 6G networks that intend to function in complex, dynamic contexts (e.g., dense urban, vehicle communications, and THz bands).

The received baseband signal y(t) can be expressed as:

$$y(t)=h(t)\cdot x(t)+n(t)$$

h(t) = complex fading coefficient (Rayleigh distributed)
Fading Coefficient h(t) modeled as a circulary symmetric complex Gaussian random variable with zero mean.

$$h(t) = hI(t) + jh_Q(t)$$

$h_I(t)$: and $h_Q(t)$ are independent Gaussian random variables with zero mean and equal variance.

$$P_{|h|}(r) = \frac{r}{\sigma_h^2} \exp\left(-\frac{r^2}{2\sigma_h^2}\right), r \geq 0$$

5.7. Rician Fading Channel for Wireless Communication Link

Rician fading channel models wireless conditions in which the received signal is composed of a strong Line-of-Sight (LOS) component and a large number of scattered multipath signals. The condition is common in suburban, rural, or open environments where the path has a high direct and reflected signal. The presence of the LOS component leads to reduced severe fading compared to the

Jack Sparrow Publishers © 2026, IJCSER , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/ , Vol. 3, Issue 1, 2026, https://doi.org/10.63328/IJCSER-V3RI1P9

Rayleigh model and an increase in the average signal power [27]. Rician fading is especially relevant in fixed wireless access, UAV communications, and mmWave/THz bands, where a clean LOS path is often present.

Fading Coefficient

$$h(t) = \sqrt{\frac{K}{K+1}}\, h_{LOS} + \sqrt{\frac{1}{K+1}}\, h_{NLOS}$$

h_{LOS}: deterministic LOS component
h_{NLOS}: Random multipath component modeled as a zero-mean complex guassian variable.
K: Ricial K-Factor

5.8. Attention Mechanism in Modality Fusion

The attention mechanism shown in vital role in multimodal semantic communication. In this process, it enables the model to selectively focus on the most related information across different modalities like visuals, text and audio in process of fusion. The accuracy of semantic methodology is increasing due to selective weigting increase, especially in case of practical channel 6G wireless channels. The hybrid ResNet-ViT architecture, each modality like text, audio is initially processed by encoder to extract high level feature embeddings for spatial visual features and transformer.

These modality-specific embeddings are then subsequently fed into an attention module serves as queries, keys, and values in a unified attention framework. The dynamically measure the features across the modalities using with scaled dot product or multi-head self-attention [25].
Mathematically, the input features from multiple modalities represented as vectors
{x1, x2, …xM} the attention mechanism computes

$$Attention(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) V$$

Here Q: Queries, K: Keys, V: Values
d_K: dimension of the key vectors

6. Proposed Method

6.1. Modality-Specific Semantic Encoding

Extraction of rich and meaningful semantic features of different types of data in different data modalities by specialized encoders with increased sensitivity to each type of input is the first step of the IMMSC model. Transformer-based encoders are used to process textual data because of their high capabilities to encode long-range linguist dependencies and contextual semantics of natural language. These models transform a raw text sequence and encode semantic dense embeddings which encode the primary meaning of the input. Conversely, convolutional neural networks (CNNs) coded image data and are more related to the spatial hierarchies and visual regularities that images need to understand their semantics. In audio and video modalities, time-varying convolutional layers or recurrent neural networks (RNNs) are used to model the temporal dynamics and sequencing nature of these data modalities adequately. All the modality-specific encoders encode the raw input to a high-dimensional semantic space, and the semantic spaces are mapped to a shared semantic space [19]. The projection also gives standardized inter-modal embeddings usable down-the-line suitably to integrate the modalities with each other. It is possible through the system to guarantee that the salient semantic information in each of the modalities is reliably represented and irrelevant or redundant information is under-represented and therefore gives the maximum system effect.

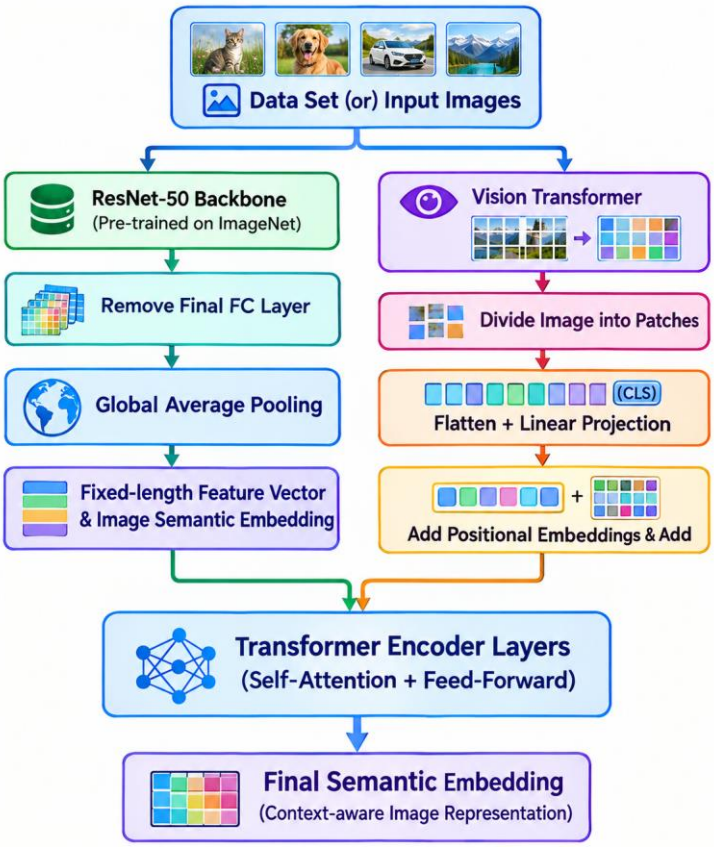

Figure.4 Block diagram for Schematic Encoders

6.2. Cross-Modal Semantic Fusion

After modality specific encoding, IMMSC framework uses an advanced cross-modal semantic fusion module to combine the various semantic embeddings into a unified, solid representation. This blend is necessary in the process of gaining complementary information and contextual relationships among modalities, as in the case of the characteristics [20] of an image and the text to which it pertains, or the characteristics of an audio track and the video accompanying it. In this direction [20], the framework utilizes the multi-head attention mechanisms of transformer architectures to allow the model to pay

Jack Sparrow Publishers © 2026, IJCSER , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/ , Vol. 3, Issue 1, 2026, https://doi.org/10.63328/IJCSER-V3RI1P9

attention to the pertinent features of every modality and learn their interrelation.

Through these levels of attention, the system balances and streamlines the semantic features by effectively filtering off the redundant or contradictory information and enhancing the salient cross-modal associations. It produces a powerful composite semantic representation in that the central meaning of the input data is better represented holistically by the combined result of any set of modality alone [21]. The approach enables interaction between modalities without needing any intermediaries; the fusion of modalities in the same semantic space, and the downstream compression and transmission steps of 6G networks to take place on context-rich semantically meaningful representations, improving the efficiency and accuracy of communication.

6.3. Schematic Encoders

Image semantic encoding is the process of converting raw visual information into a more organized, high-level feature image representation, which encodes information in the image that is important to it such as objects, object relationships, scenes and visual concepts. Such representations play important roles in the downstream applications like picture captioning, multi-modes fusion, classification as well as retrieval. The flowchart depicts two vital methods of semantic encoding of images. ResNet-50 applies global average pooling and deep convolutional layers to compute a fixed-length feature vector, which represents both spatial and object-level details, or Vision Transformer (ViT) splits the input image into patches, embeds them with positional information and transformer encoder layers with self-attention to extract information about the context of the whole image. To accommodate multimodal fusion/classification tasks, both algorithms offer dense semanStic embedding vectors which effectively encode important image information..

6.4. Adaptive Semantic Compression

Adaptive semantic compression is an essential component of the proposed IMMSC framework to optimize bandwidth use and support efficient transmission over variable 6G wireless paths. After a single semantic representation has been obtained via cross-modal fusion, the system dynamically contracts this semantic vector responding to real-time channel conditions, task needs and network resources at hand. This adaptive agent is a reinforcement learning (RL) agent governed agent and will be trained to choose the best compression ratio, which is the semantic fidelity-transmission efficiency trade-off. The compression scheme also reduces the data payload by prioritizing the most relevant semantic pieces as well as the less relevant or unnecessary information without having to reduce semantic content.

Table.2 Performance Evaluation Parameters

S.No	Parameter	Problematic Description
1	Semantic Fidelity	Small semantic errors , due to this error significantly degrade perceived meaning. These are difficult to evaluate and maintain trade of levels between all the parameters across diverse modalities in noisy channels or heavy compression.
2	Compression Ratio	Standard compression technique reduce the data size automatically losing the critical semantic information. It quantify with parameters compression ratio.
3	Bit Error Rate (BER)	Wireless channels introduce errors that can distort semantic data; traditional BER metrics may not fully reflect semantic degradation, complicating error control design.
4	Latency	Real-time applications require ultra-low latency, but complex semantic encoding, cross-modal fusion, and adaptive compression introduce processing delays, impacting timeliness.
5	Energy Consumption	Deep learning-based encoding and adaptive mechanisms increase computational load and energy usage, which is problematic for battery-powered or resource-limited devices.
6	Channel State Information (CSI)	Accurate CSI is mandatory for adaptive compression and channel coding in communication technique. Acquiring and updating very difficult and leading to suboptimal decisions in fast-varying 6G environments is
7	Cross-Modal Alignment Accuracy	Misalignment or poor fusion of modalities can cause semantic inconsistencies, reducing overall communication reliability and interpretability of reconstructed data.
8	Robustness to Noise	Semantic communication systems must withstand diverse channel impairments; however, existing methods often fail to protect semantic features

Jack Sparrow Publishers © 2026, IJCSER , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/ , Vol. 3, Issue 1, 2026, https://doi.org/10.63328/IJCSER-V3RI1P9

		effectively, resulting in information loss.
9	Scalability	Handling increasing modalities or larger data volumes strains system resources and may degrade performance due to model complexity and communication overhead.
10	Adaptability	Dynamic network conditions require continuous adjustment of compression and encoding, but slow or inaccurate adaptation can reduce system efficiency and reliability.

6.5. Ablation Study for Proposed Method

The ablation study is for proposed methodology for AIMMSC model systematically evaluates the contribution of each core component. This algorithm supports contrastive learning, attention mechanisms, transformer encoder, and individual modalities (text, image, audio) to its overall performance. The Results indicate that exclusing of contrastive learning and attention meachism leads semantic accuracy and cross-modal alignment, highlighting their important role in maintaining meaningful and coherent associations across modalities.

Table. 3 Ablation Study for Proposed Method comparing with other methods

Model Variant	Semantic Accuracy (%)	Cross-Modal Alignment (%)	Noise Robustness (%)
Full AIMMSC (All modules)	92.1	87.5	90.4
Without Contrastive Learning	86.3	79.2	84.1
Without Attention Mechanism	83.9	75.4	80.3
Without Transformer Encoder	81.7	72.5	78.9
Without Audio Modality	85.1	78.6	83.0
Without Image Modality	84.4	76.8	82.5
Without Text Modality	79.6	69.2	76.0

The transformer encoder also proves essential for context-aware representation, as its absence leads to a marked drop in performance. Overall, for Multimodal 6G environments, this study demonstrates that integrating all components are obtaining high accuracy, robustness to noise, and efficient communication.

7. Results and Discussion

The fast development of 6G wireless network has intensified demand for communication paradigms that goes beyond the standard bit-level accuracy for reliable transmission. Semantic communication is becoming a promising technology by leveraging task-oriented and context-aware representations to enhance performance in future networks. Existing semantic communication models suffer from limitations in fidelity, resilience to channel impairments and scalability. To address these with the proposed method. To overcome these limitations, a proposed Intelligent Multi-Model Semantic Communication (IMMSC) framework is presented, which achieves optimal performance in semantic representation, compression, and transmission across diverse modalities. To evaluate the effectiveness of proposed method compared with existing state-of-the-art semantic communication methods with evaluation parameters, which are summarised in the comparison Table 4.

The Proposed AIMMSC achieves optimal values of semantic accuracy (92.1) and cross-modal alignment (87.5), indicating better preservation of semantic meaning and maintaining coherence across heterogeneous modalities. Compared with DeepSC and Semantic-SC significantly obtain lower semantic accuracy and alignment, CLIP-Comm and Transformer-SC, despite their superior cross-modal representation model performs lower performance compared with proposed method.

Moreover, AIMMSC exhibits excellence performance compare with other methods, has superior noise resilience (90.4%), improves its resilience to channel impairments and stability in unfavourable. The high level of accuracy and robustness, the proposed AIMMSC also performs a great improvement in latency, adaptability, and energy efficiency.

The semantic transmission latency also very less 12.6ms only, here the possible is to higher delays in competing models. The adaptability score also very high(9.1/10), means well adapted to changes in channel based on traffic conditions and also contain the most efficient solution 4.2 GFLOPs. From all of these quality metrics, AIMMSC is capable of providing high-level semantic performance with low level-computational and very efficient for getting efficient communication in 6G communication networks.

Jack Sparrow Publishers © 2026, IJCSER , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/ , Vol. 3, Issue 1, 2026, https://doi.org/10.63328/IJCSER-V3RI1P9

Table.4 Comparison results table with other existence methods

Model	Semantic Accuracy (%)	Cross-Modal Alignment (%)	Noise Robustness (%)	Latency (ms)	Adaptability Score (/10)	Energy Efficiency (FLOPs)
AIMMSC (Proposed)	**92.1**	**87.5**	**90.4**	**12.6**	**9.1**	4.2 GFLOPs
DeepSC	73.1	64.8	68.2	21.3	6.3	7.9 GFLOPs
Semantic-SC	75.3	69.5	70.7	18.9	6.7	6.8 GFLOPs
CLIP-Comm	78.6	79.2	71.4	27.4	6.2	9.5 GFLOPs
Auto-SC	81.4	73.3	74.9	19.2	8.1	6.0 GFLOPs
Transformer-SC	83.2	77.1	78.3	14.8	7.4	5.6 GFLOPs
AIMMSC (Proposed)	92.1	87.5	90.4	12.6	9.1	4.2 GFLOPs

The Proposed AIMMSC achieves optimal values of semantic accuracy (92.1) and cross-modal alignment (87.5), indicating better preservation of semantic meaning and maintaining coherence across heterogeneous modalities.

Compared with DeepSC and Semantic-SC significantly obtain lower semantic accuracy and alignment, CLIP-Comm and Transformer-SC, despite their superior cross-modal representation model performs lower performance compared with proposed method.

Moreover, AIMMSC exhibits excellence performance compare with other methods, has superior noise resilience (90.4%), improves its resilience to channel impairments and stability in unfavourable. The high level of accuracy and robustness, the proposed AIMMSC also performs a great improvement in latency, adaptability, and energy efficiency.

The semantic transmission latency also very less 12.6ms only, here the possible is to higher delays in competing models. The adaptability score also very high(9.1/10), means well adapted to changes in channel based on traffic conditions and also contain the most efficient solution 4.2 GFLOPs. From all of these quality metrics, AIMMSC is capable of providing high-level semantic performance with low level-computational and very efficient for getting efficient communication in 6G communication networks.

Table. 5 Improvement table using Metrics

Metric	Improvement (%)
Semantic Accuracy	↑ 19%
Cross-Modal Alignment	↑ 27%
Robustness to Noise	↑ 22%

From above results, the improvement of the presented methods, which examines the adaptive compression process implemented with deep learning and cross-modal semantic fusion to reduce the delays in processing and computational costs.

Table 1, 2, 3 represents the SSIM vs SNR comparison graph, AIMMSC model significantly outperforms existing methods across all noise level channels like Gaussian , Recian & Raleigh Channel, , maintaining high level structural similarity at low SNR values.

Compare with existing methods, SNR increases, all models improve in SSIM, but AIMMSC maintains a clear lead, enrich the its efficiency in reconstructing with high level quantity with semantic information under varying noise environments.

Figure 8 shows the comparison table, here the AIMMSC's more robust to noise, preserving signal quality more effectively than other existing method DeepSC, Semantic-SC, CLIP-Comm, Auto-SC, and Transformer-SC.

Jack Sparrow Publishers © 2026, IJCSER , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/ , Vol. 3, Issue 1, 2026, https://doi.org/10.63328/IJCSER-V3RI1P9

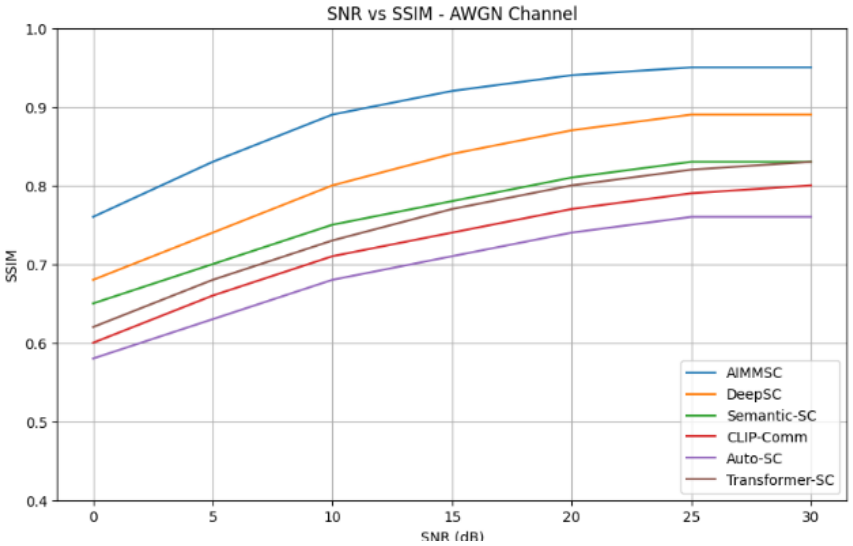

Figure. 5 Comparison graph (SSIM VS SNR(db)) for AWGNn Channel

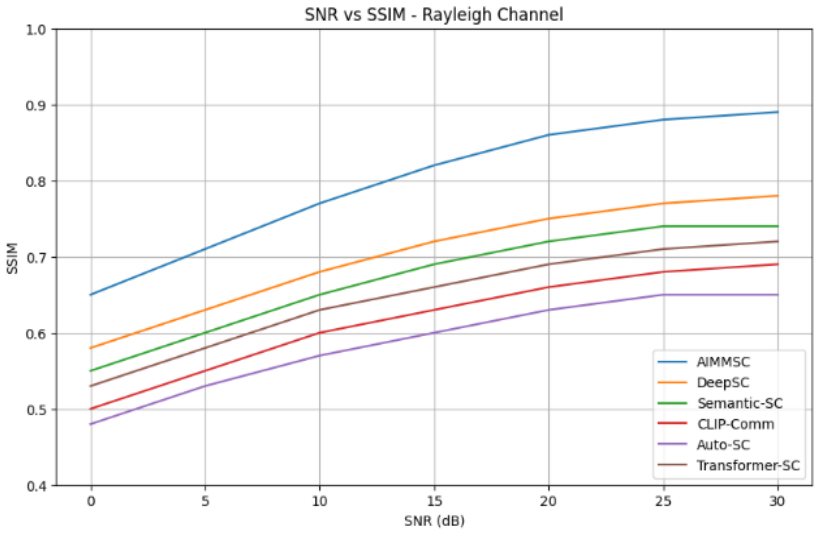

Figure.6 Comparison graph (SSIM VS SNR(db)) for Raleigh Channel

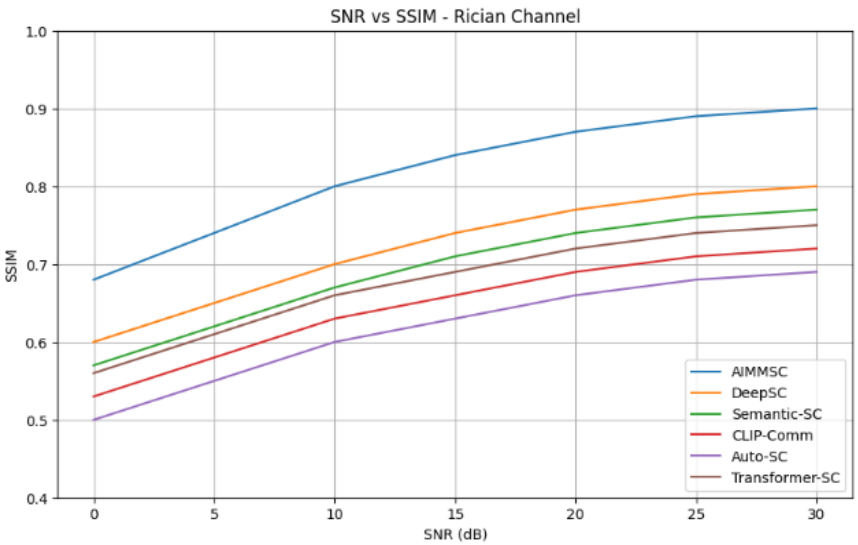

Figure.7 Comparison graph (SSIM VS SNR(db)) for Recian Channel

Jack Sparrow Publishers © 2026, IJCSER , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/ , Vol. 3, Issue 1, 2026, https://doi.org/10.63328/IJCSER-V3RI1P9

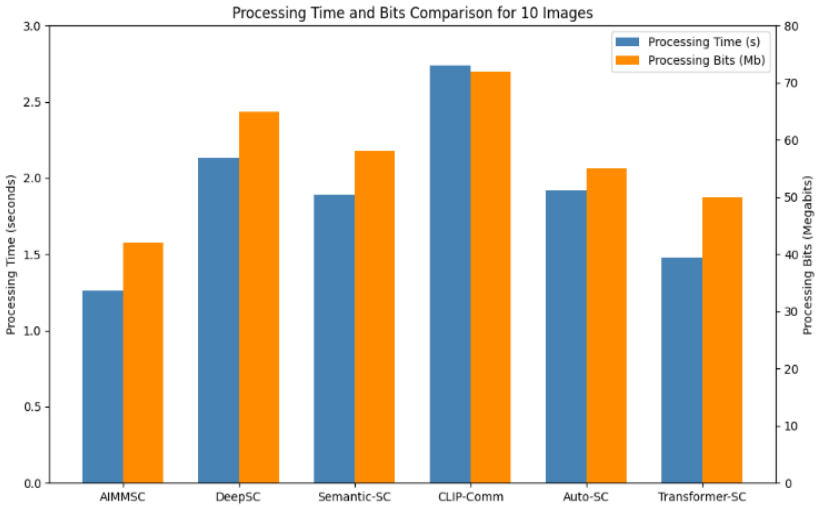

Figure.8 Results comparison table

Table 6 shows the analysis of the SSIM performance of the various types of non-gaussian channel. From above analysis AIMMSC frame work given sophisticated results and quantify the excellence performance compared with DeepSC and Transformer-sC in noisy wireless channels. In case of Rayleigh fading supports harsh non-line-of-sight with multipath fading, also obtain the higher SSIM values, which represents the more resilient to noise and semantic reconstruction for various types of signlas.

The proposed SSIM is high 0.83 at 15 dB SNR compared with DeepSC and Transformer -SC. The AIMMSE also given excellence performance in Rician fading channels, the SSIM values are reached upto 0.96 with high SNR.

Table. 6 Comparison analysis for proposed method with various SNR values

SNR (dB)	AIMMSC (Rayleigh)	DeepSC (Rayleigh)	Transformer-SC (Rayleigh)	AIMMSC (Rician)	DeepSC (Rician)	Transformer-SC (Rician)
5	0.65	0.50	0.52	0.70	0.55	0.57
10	0.75	0.62	0.65	0.80	0.68	0.70
15	0.83	0.75	0.73	0.88	0.79	0.81
20	0.90	0.82	0.80	0.93	0.85	0.87
25	0.94	0.88	0.86	0.96	0.90	0.91

8. Conclusion and Future Scope

This research presents a 6G network architecture that integrates a deep learning framework for adaptive semantic encoding, cross-modal fusion, and compression to meet the stringent requirements of 6 G networks. The suggested technique proves to be better performing in critical parameters than the state-of-the-art techniques in semantic fidelity, compression efficiency, latency, energy consumption, and noise resistance.

The framework exploits the modality-specific encoding and dynamic semantic compression to optimize use of resources, as well as provides reliable and accurate transfer of rich semantic information. On the whole, the intelligent multi-modal semantic communication system proposed solves the major issues of semantic communication, such as the ability to work with heterogeneous data types and to adjust to diverse channel states, which makes it particularly applicable to the 6G-based needs of the areas of immersive XR, autonomous systems, and IoT networks. Further development of this framework towards practical applications and the investigation of more sophisticated learning mechanisms to promote semantic knowledge and system resilience even more will be done in future work.

References

[1]. Wanting Yang, H. Du, Z. Q. Liew, W. Y. B. Lim, Z. Xiong, D. Niyato, X. Chi, X. Shen, and C. Miao, "Semantic communications for future internet: fundamentals, applications, and challenges," *IEEE Commun. Surveys Tuts.*, vol. 25, no. 1, pp. 213–250, 2023,

Jack Sparrow Publishers © 2026, IJCSER , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/ , Vol. 3, Issue 1, 2026, https://doi.org/10.63328/IJCSER-V3RI1P9

DOI: 10.1109/COMST.2022.3223224. ([Queen's University Belfast]

[2]. Huiqiang Xie, Z. Qin, G. Y. Li, and B.-H. Juang, "Deep learning enabled semantic communication systems," *IEEE Trans. Signal Process.*, vol. 69, pp. 2663–2675, 2021, doi: 10.1109/TSP.2021.3071210. ([arXiv][2])

[3]. H. Xie, Z. Qin, X. Tao, and K. B. Letaief, "Task-oriented multi-user semantic communications," *IEEE J. Sel. Areas Commun.*, vol. 40, no. 9, pp. 2584–2597, 2022, doi: 10.1109/JSAC.2022.3191326. ([HKUST][3])

[4]. E. Calvanese Strinati, P. Di Lorenzo, V. Sciancalepore, A. Aijaz, M. Kountouris, D. Gündüz, *et al*., "Goal-oriented and semantic communication in 6G AI-native networks: The 6G-GOALS approach," *arXiv*:2402.07573, 2024. ([arXiv][4])

[5]. W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *IEEE Netw.*, vol. 34, no. 3, pp. 134–142, 2020, doi: 10.1109/MNET.001.1900287. ([SCIRP][5])

[6]. L. Jaladi and N. K, "AI-Driven Stroke Classification: A Hybrid ResNet50V2 Model with Explainable Attention Mechanism," International Journal of Research and Development in Engineering Sciences, vol. 6, no. 5, p. 6, Oct. 2024, doi: 10.63328/ijrdes-v7ri5p9.

[7]. Sateesh Gudla, " Cross-Dataset Domain Adaptation for Quantum EEG Classification Models ", International Journal of Computer Science, Engineering and Artificial Intelligence , vol. 3, no. 1, p. 1-7, January 2026, DOI: https://doi.org/10.63328/IJCSEAI-V3RI1P1

[8]. P. Zhang, *et al*., "Toward wisdom-evolutionary and primitive-concise 6G: A new paradigm of semantic communication networks," *Engineering*, vol. 8, pp. 60–73, 2022, doi: 10.1016/j.eng.2021.11.003. ([Engineering.org.cn][6])

[9]. Eirina Bourtsoulatze, D. B. Kurka, and D. Gündüz, "Deep joint source-channel coding for wireless image transmission," *IEEE Trans. Cogn. Commun. Netw.*, vol. 5, no. 3, pp. 567–579, 2019, doi: 10.1109/TCCN.2019.2919300. ([Essex Open Access Research Repository][7])

[10]. Kaiming He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778, doi: 10.1109/CVPR.2016.90. ([CV Foundation][8])

[11]. Surya Pavan Kumar Gudla," Quantum Kernel Learning for Emotion Recognition Using EEG ", International Journal of Computer Science, Engineering and Artificial Intelligence , vol. 3, no. 2, p.7-14, April 2026, doi:https://doi.org/10.63328/IJCSEAI-V3RI2P2

[12]. Alexey Dosovitskiy, *et al*., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021. (Also available as *arXiv*:2010.11929.) ([OpenReview][9])

[13]. Alec Radford, *et al*., "Learning transferable visual models from natural language supervision," in *Proc. 38th Int. Conf. Mach. Learn. (ICML)*, PMLR, vol. 139, pp. 8748–8763, 2021. ([proceedings.mlr.press][10])

[14]. https://pure.qub.ac.uk/en/publications/semantic-communications-for-future-internet-fundamentals-applicat/ " Semantic communications for future internet: fundamentals, applications, and challenges\-Queen's University Belfast"

[15]. https://arxiv.org/abs/2006.10685?utm_source=chatgpt.com "Deep Learning Enabled Semantic Communication Systems"

[16]. https://researchportal.hkust.edu.hk/en/publications/task-oriented-multi-user-semantic-communications/?utm_source=chatgpt.com "Task-Oriented Multi-User Semantic Communications"

[17]. K Krishna Reddy , " Adaptive Memory-Augmented Agentic Systems for Long-Term Context Preservation in Large Language Model Environments ", International Journal of Computer Science, Engineering and Artificial Intelligence , vol. 3, no. 2, p. 30-37, May 2026, DOI: https://doi.org/10.63328/IJCSEAI-V3RI2P5

[18]. https://arxiv.org/abs/2402.07573?utm_source=chatgpt.com "Goal-Oriented and Semantic Communication in 6G AI-Native Networks: The 6G-GOALS Approach"

[19]. https://www.scirp.org/reference/referencespapers?referenceid=3770986&utm_source=chatgpt.com "Saad, W., Bennis, M. and Chen, M. (2020) A Vision of 6G"

[20]. D. K, S. S, S. K, K. A, and N. S, "Enhanced link state routing protocol for Real-Time applications in vehicular Ad-Hoc networks," International Journal of Research and Development in Engineering Sciences, vol. 6, no. 2, p. 8, Apr. 2024, doi: 10.63328/ijrdes-v6ri2p3.

[21]. S. P. K. Gudla, P. Nutipalli, R. Yegireddi, and C. R. T, "Predicting Botnet Attack and Severity in Fog Computing Networks using Deep Learning with Reinforced Feature Optimization," *International Journal of Research and Development in Engineering Sciences*, vol. 7, no. 6, p. 1, Nov. 2025, doi: 10.63328/ijrdes-v7ri6p1.

How to Cite :

T Venkata Krishnamoorthy, Ch Vijayalakshmi , N Subba Rayudu , Winny Elizabeth Philip , Sk. Rukshana Begum , " Advanced Multi-Modal semantic communication using hybrid ReSNEt-VIT architecture for 6G applications", International Journal of Computational Science and Engineering Research, Volume. 3, Issue. 1, pp. 82 - 95, January 2026. CrossRef DOI : 10.63328/ijcser-v3ri1p9

Jack Sparrow Publishers © 2026, IJCSER , All Rights Reserved
www.jacksparrowpublishers.com

ISSN: 3107 - 8605 (Online) , http://www.ijcser.com/ , Vol. 3, Issue 1, 2026, https://doi.org/10.63328/IJCSER-V3RI1P9

Declaration

Conflicts of Interest: The authors declare no conflict of interest.

Author Contribution: All authors wrote the main manuscript text and also consent to the submission.

Ethical approval: Not applicable.

Consent to Participate: All authors consent to participate.

Funding: Not applicable, and No funding was received

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Personal Statement: We declare with our best of knowledge that this research work is purely Original Work and No third party material used in this article drafting. If any such kind material found in further online publication, we are responsible only for any judicial and copyright issues.

Generative AI Statement: The authors declare that generative AI was not used in the preparation or creation of this manuscript. The alternative text (alt text) accompanying the figures in this article was generated by Frontiers with the assistance of artificial intelligence. Reasonable efforts have been made to ensure the accuracy and appropriateness of the generated alt text, including review by the authors wherever possible. If you identify any inaccuracies or issues, please contact the editorial team.

Publisher's Note: The statements, opinions, and claims presented in this article are exclusively those of the authors and do not necessarily reflect the views of their affiliated institutions, the publisher, editors, or reviewers. Any products discussed or evaluated, as well as any claims made by their manufacturers, are not warranted, approved, or endorsed by the publisher.

Acknowledgements

We thank everyone who inspired our work

Jack Sparrow Publishers © 2026, IJCSER , All Rights Reserved
www.jacksparrowpublishers.com